1-1-1980

Calculation of Probability of Correct Classification for Two-Class Gaussian Classifiers with Arbitrary Hyperquadratic Decision Boundaries

Arthur G. Wacker

Talaat S. El-Sheikh

Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

and

Soil Information Systems

and

Remote Sensing and Soil Survey

June 3-6, 1980

Proceedings

The Laboratory for Applications of Remote Sensing

Purdue University
West Lafayette
Indiana 47907 USA

IEEE Catalog No.
80CH1533-9 MPRSD

CALCULATION OF PROBABILITY OF CORRECT CLASSIFICATION FOR TWO-CLASS GAUSSIAN CLASSIFIERS WITH ARBITRARY HYPERQUADRATIC DECISION BOUNDARIES

ARTHUR G. WACKER AND TALAAT SALEM EL-SHEIKH

University of Saskatchewan
Canada

I. ABSTRACT

In this paper a technique is developed in order to numerically calculate the hypervolume under a multidimensional Gaussian function over a region of the space defined by an arbitrary hyperquadratic boundary. The technique is a modified version of the technique developed by Fukunaga and Krile.[2] The latter technique can be used only if the hyperquadratic boundary results from the intersection of the Gaussian function, under which the hypervolume is being calculated, with some other Gaussian function as opposed to an arbitrary hyperquadratic boundary.

A practical problem in which the hypervolume calculation mentioned above is of interest arises in statistical pattern classification involving Gaussian classes. In this situation the Gaussian function under consideration is actually a probability density function and the arbitrary hyperquadratic boundary results from the intersection between two estimated distributions which partitions the feature space into two disjoint decision regions. For this case the hypervolume under the probability density function of any class in the region for which patterns are classified into that class, is actually the probability of correctly classifying vectors from the class.

The proposed technique has been successfully implemented and it has proven to be quite efficient and reasonably simple. Real data have been used to demonstrate the applicability and efficiency of the technique and to study the effect of estimation on the value of the probability of correct classification.

II. INTRODUCTION

In statistical pattern classification it is assumed that measurement vectors representing the patterns from any particular class are random vectors originating from some multivariate distribution. In many situations a desirable objective of a classification rule is to maximize the probability of correct classification. If the statistical approach is utilized and it is assumed that the class-conditional distributions and prior class probabilities are known to the classifier, then it is logical to use Bayes' rule[1] which is known to achieve the stated objective.

In "real life" pattern classification problems the true class distributions are naturally never known to the classification system designer but a number of training patterns might be available from each class. Under these circumstances, a commonly used approach is to again use Bayes' rule with the true class distributions replaced by their estimates. In this paper the class distributions are assumed to be Gaussian. For this case, in general, the decision boundaries that result from the above rule are hyperquadratic.

For our purpose it is essential to clearly distinguish between the unknown true (underlying) class-conditional distributions from which patterns are assumed to arise and the estimated distributions that are used in the classification rule. To distinguish between these distributions the terms true distributions and estimated distributions respectively will be consistently used.

In situations where maximizing the probability of correct classification is the objective it is natural to use overall probability of correct classification as a performance indicator for the classifier. This probability can be obtained by calculating the probability of correctly classifying vectors from each class and then summing up these probabilities after weighting each by its prior probability. The class-conditional probability of correct classification for any class is equal to the hypervolume under the true class-conditional probability density function of this class over the region for which patterns are classified into this class. This region is delineated by boundaries established by the classification rule. In the modified Bayes' rule previously described these boundaries are established by the estimated distributions.

Obviously the volumes of interest can't be calculated in any "real life" classification problem since the true density functions are not available. Consequently, the only possible way of

CH1533-9/80/0000-0294 $00.75

obtaining some index of classifier performance is to estimate each of these volumes. In practice, this has typically been done by either classifying additional test vectors from the true distributions or by calculating the hypervolumes under the estimated probability density functions. This latter approach has been used by Fukunaga and Krile[2] and Hallum[3] both for two-class multivariate Gaussian problems and by Mobasseri and McGillem[4] for multiclass Gaussian problems.

It is of considerable interest to determine the effect on the probability of correctly classifying vectors from the true distributions caused by using estimated rather than true distributions in the classification rule. This problem can only be studied theoretically since, as already noted, in any "real life" classification problem the true distributions are naturally not available. For theoretical studies the true distributions are assumed to be completely specified but in order to simulate the "real life" situation it is assumed that these distributions can't be used by the classifier. Consequently, as in the "real life" case, a number of training vectors generated according to each true class distribution are available to estimate the distribution for that class. The estimated distributions are then used, identical to the "real life" situation, in the classification rule. In this theoretical framework, since the true distributions are available, it is possible (though it might be difficult) to calculate the hypervolume under each true class distribution over the decision region associated with the corresponding class.

For two-class Gaussian problems Fukunaga and Krile have developed a technique to very accurately calculate the hypervolume under any Gaussian distribution within a region defined by a specific hyperquadratic decision boundary. The specific hyperquadratic decision boundary results from the intersection of the Gaussian distribution, under which the hypervolume is being calculated, with any other Gaussian density function. If the hyperquadratic decision boundary results from the intersection of two Gaussian density functions both different from the class distribution under which the hypervolume is being calculated (i.e. estimated distributions), then Fukunaga's technique can't be directly used. This paper describes how Fukunaga's technique can be modified in order to numerically calculate the required hypervolume. Utilizing this modification, the true overall probability of correct classification of a multivariate two-class Gaussian problem can be calculated directly as opposed to estimating this quantity using the approaches previously mentioned.

In section III the proposed technique is mathematically developed for the two-class Gaussian problem. In section IV a "real life" pattern classification problem is considered in some detail.

III. MATHEMATICAL DEVELOPMENT

For the two-class problem it is assumed that patterns can arise from one of two classes C_1 and C_2 of prior probabilities P_1 and P_2. The two class-conditional probability density functions are assumed to be multivariate Gaussian with mean vectors M_1 and M_2 and covariance matrices Σ_1 and Σ_2. Since in any "real life" pattern classification problem all the parameters are naturally unknown, it is assumed here that the above parameters including the priors are not available to the classifier. It is also assumed that a number of training (design) vectors are available from each class.

Using the available design vectors, all the unknown parameters are estimated and these estimates are substituted into Bayes' classification rule in place of the unknown true parameters. Assuming that $\hat{P}_i$, $\hat{M}_i$ and $\hat{\Sigma}_i$ are the estimates for class C_i, the following classification rule results

$$(X-\hat{M}_1)^T\hat{\Sigma}_1^{-1}(X-\hat{M}_1) - (X-\hat{M}_2)^T\hat{\Sigma}_2^{-1}(X-\hat{M}_2)$$

$$+ \ell n[|\hat{\Sigma}_1|/|\hat{\Sigma}_2|] - 2\ell n[\hat{P}_1/\hat{P}_2] \lesssim 0 \rightarrow X\varepsilon\{^{C_1}_{C_2} \quad (1)$$

where X is an observation vector from either C_1 or C_2. This rule, in general, produces a hyperquadratic decision boundary between the two classes. This is true even for the special case in which the true distributions have equal covariance matrices (i.e. $\Sigma_1=\Sigma_2$) unless this information is known to the classifier a priori.

It is required to calculate the overall probability of correct classification P_{cr} resulting from using rule (1). This probability can be written as follows:

$$P_{cr} = \sum_{i=1}^{2} P_i P_{cr}^{(i)} , \quad (2)$$

where $P_{cr}^{(i)}$ is the probability of correctly classifying vectors from the true distribution of class C_i using rule (1). This probability is equal to the hypervolume under the true probability density function of class C_i over the region where vectors are classified into this class by rule (1). As previously mentioned, it is not possible to calculate this hypervolume in any "real life" pattern classification problem due to lack of knowledge of the true densities. In fact it is very difficult to calculate the hypervolume even in theoretical studies (in which the true densities can be assumed to be available) since direct calculation involves a multivariate integration over a region with complicated boundaries.

In order to simplify the analysis a new univariate random variable is introduced as

$$W(X) = (X-\hat{M}_1)^T\hat{\Sigma}_1^{-1}(X-\hat{M}_1) - (X-\hat{M}_2)^T\hat{\Sigma}_2^{-1}(X-\hat{M}_2)$$

$$+ \ell n[|\hat{\Sigma}_1|/|\hat{\Sigma}_2|] - 2\ell n[\hat{P}_1/\hat{P}_2], \quad (3)$$

where X, as defined before, is an observation vector from either C_1 or C_2. Consequently rule (1) reduces to

$$W(X) \lesseqgtr 0 \rightarrow X\varepsilon\{\begin{matrix}C_1\\C_2\end{matrix} . \qquad (4)$$

In the course of the paper the two univariate random variables W_1 and W_2 will be used where

$$W_i \triangleq W(X|C_i), \quad i=1 \text{ or } 2,$$

that is, W_i is the statistic W(X) given that observations arise from class C_i.

The two univariate random variables W_1 and W_2, introduced above, contain all the information about classes C_1 and C_2 that is required for classification purposes. Assuming that the density functions of these two random variables can be derived, $P_{cr}^{(1)}$ and $P_{cr}^{(2)}$ can be respectively determined by integrating the density function of W_1 between $-\infty$ and zero and the density function of W_2 between zero and $+\infty$. Thus each of the complicated multivariate integrals, mentioned before, is now replaced by a simple univariate integration which can be easily evaluated numerically.

Since deriving the density function of W_1 or W_2 when classification is based on estimated distributions is generally difficult, simplifying assumptions have been in the past considered. Thus observations have been assumed to have come from the estimated distribution (as opposed to the true distribution) of either C_1 or C_2. In other words, the probability of correctly classifying vectors from the estimated distributions is calculated and used as an estimate for the true value of this probability. For this approach, a technique has been developed by Fukunaga and Krile[2] to numerically calculate this quantity. This same technique can be also used to calculate the true probability if the true parameters are known to the classifier. Since the technique developed here is presented as an extension of Fukunaga's technique, this technique is briefly described.

The first step in Fukunaga's technique is to simultaneously diagonalize the two estimated matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ by applying a nonsingular linear transformation to the observations. This transformation substantially simplifies the calculations and in the same time it does not affect how patterns are classified by rule (1) or (4) and consequently does not affect the value of P_{cr}.

The direct derivation of the probability density function of W_1 or W_2 is very difficult. Fukunaga's technique avoids the direct derivation of the densities of W_1 and W_2 by utilizing instead their more easily derived characteristic functions. To calculate the probability of correct classification a basic transform theorem for integrals enables the necessary areas under the densities of W_1 and W_2 to be calculated by performing a suitable integration of the characteristic functions. By this approach the probability of correct classification can be expressed in terms of two univariate integrals of reasonably well behaved functions. The above steps will be further clarified in the course of development of the technique proposed in this paper.

In this paper it is required to calculate the true probabilities of correct classification, i.e. the hypervolumes under the true densities. In other words in rule (1) or (4) X is assumed to arise from one of the true distributions. Thus simplifying W_1 (or W_2) might seem to require the diagonalization of $\hat{\Sigma}_1$, $\hat{\Sigma}_2$ and Σ_1 (or Σ_2) which is now the covariance matrix of X. This diagonalization is not possible since only two matrices can be simultaneously diagonalized. In order to resolve this difficulty, W(X) is first written in the following form.

$$W(X) = X^T(\hat{\Sigma}_1^{-1}-\hat{\Sigma}_2^{-1})X - 2(\hat{M}_1^T\hat{\Sigma}_1^{-1}-\hat{M}_2^T\hat{\Sigma}_2^{-1})X$$

$$+ \{\hat{M}_1^T\hat{\Sigma}_1^{-1}\hat{M}_1-\hat{M}_2^T\hat{\Sigma}_2^{-1}\hat{M}_2+\ell n[|\hat{\Sigma}_1|/|\hat{\Sigma}_2|] - 2\ell n[\hat{P}_1/\hat{P}_2]\}$$

$$= X^TAX - 2b^TX + d, \qquad (5)$$

where X is from the true distribution of either C_1 or C_2 with

$$A = \hat{\Sigma}_1^{-1}-\hat{\Sigma}_2^{-1} , \qquad (6a)$$

$$b = \hat{\Sigma}_1^{-1}\hat{M}_1-\hat{\Sigma}_2^{-1}\hat{M}_2, \qquad (6b)$$

$$d = \hat{M}_1^T\hat{\Sigma}_1^{-1}\hat{M}_1-\hat{M}_2^T\hat{\Sigma}_2^{-1}\hat{M}_2 +\ell n[|\hat{\Sigma}_1|/|\hat{\Sigma}_2|]$$

$$-2\ell n[\hat{P}_1/\hat{P}_2]. \qquad (6c)$$

In the following analysis only class C_1 is considered since class C_2 can be similarily considered. From (5) it is obvious that simplifying $W_1(X)$ in its functional form requires diagonalizing only two matrices; Σ_1 and A. Since neither $\hat{\Sigma}_1$ nor $\hat{\Sigma}_2$ needs to be diagonalized, Σ_1 and A (i.e. $\hat{\Sigma}_1^{-1}$ - $\hat{\Sigma}_2^{-1}$) are simultaneously diagonalized such that

$$\phi_1^T\Sigma_1\phi_1 = I \quad \text{and} \quad \phi_1^TA\phi_1 = \Lambda^{(1)}, \qquad (7)$$

where $\Lambda^{(1)}$ is a diagonal matrix whose diagonal elements $\lambda_1^{(1)}$, $\lambda_2^{(1)}$, ..., $\lambda_N^{(1)}$ are the eigenvalues of the matrix $(\Sigma_1^{-1}A)$ and the ith column of ϕ is an eigenvector corresponding to $\lambda_i^{(1)}$, normalized to make $\phi_1^T\Sigma_1\phi_1$ a unit matrix. In order to further simplify the calculations, a transformation is first applied such that the origin is at M_1. Thus as a consequence of these two transformations a new variable Y arises where

$$Y = \phi_1^T(X-M_1). \qquad (8)$$

Also b and d of (6) are likewise transformed by the same transformation into $E^{(1)}$ and $d^{(1)}$ as follows

$$E^{(1)} = \phi_1^T[\hat{\Sigma}_1^{-1}(\hat{M}_1-M_1) - \hat{\Sigma}_2^{-1}(\hat{M}_2-M_1)], \quad (9a)$$

$$d^{(1)} = (\hat{M}_1-M_1)^T\hat{\Sigma}_1^{-1}(\hat{M}_1-M_1) - (\hat{M}_2-M_1)^T\hat{\Sigma}_2^{-1}(\hat{M}_2-M_1)$$

$$- \ln[|\Lambda_1|] - 2\ln[\hat{P}_1/\hat{P}_2]. \quad (9b)$$

Thus substituting into (5) with $X\epsilon C_1$, W_1 can be written as a function of Y as

$$W_1(Y) = \sum_{i=1}^{N} \{\lambda_i^{(1)} y_i^2 - 2E_i^{(1)} y_i\} + d^{(1)} \quad (10)$$

where $\lambda_i^{(1)}$ and $E_i^{(1)}$ are the ith components of $\Lambda^{(1)}$ and $E^{(1)}$ respectively.

Thus $W_1(Y)$ is now expressed in terms of independent Gaussian random variables y_i each of which has zero mean and unit variance for $Y\epsilon C_1$. Following Fukunaga's approach the characteristic function of $W_1(Y)$ in (10) is obtained and consequently $P_{cr}^{(1)}$ is expressed in the form of a univariate integral of this characteristic function.[2]

The same steps can be exactly repeated for class C_2 and $P_{cr}^{(2)}$ can be expressed in the same way as $P_{cr}^{(1)}$. Thus $P_{cr}^{(i)}$, i = 1 or 2, can be written as follows

$$P_{cr}^{(i)} = \frac{1}{2} + (-1)^{i+1} \frac{1}{\pi} \int_0^{\infty} \frac{F_i(\omega)}{\omega} \sin(\theta_i(\omega))\, d\omega \quad (11)$$

where

$$F_i(\omega) = [\prod_{j=1}^{N} \{1+(2\lambda_j^{(i)}\omega)^2\}]^{-1/4}$$

$$\cdot \exp-[\sum_{j=1}^{N} \frac{2(E_{ij}\omega)^2}{1+(2\lambda_j^{(i)}\omega)^2}] \quad (12a)$$

and

$$\theta(\omega) = [\sum_{j=1}^{N} \frac{\lambda_j^{(i)}(2E_{ij}\omega)^2}{1+(2\lambda_j^{(i)}\omega)^2} - d^{(i)}]\omega$$

$$- \frac{1}{2} \sum_{j=1}^{N} \tan^{-1}(2\lambda_j^{(i)}\omega). \quad (12b)$$

Using numerical integration on (11), the two probabilities of correct classification can be determined for each combination of the true and estimated parameters. As has been mentioned in [2], the number of sampling points required for convergence of the integrals in (11) are relatively small especially for higher dimensionality.

It is important to mention that the proposed technique can, in general, be used to numerically calculate the hypervolume under any Gaussian function (not necessarily a probability density function) over a region defined by an arbitrary hyperquadratic decision boundary.

IV. EXPERIMENTAL RESULTS

In order to demonstrate the applicability of the proposed technique, a simulation study was carried out on a real set of data. This set of data was the result of eight tests, where each test gives rise to one feature, performed by Marill and Green[5] on the hand-printed letters A, B, C and D. Using 200 training vectors for each letter and assuming that each group of 200 vectors had arisen from some 8-dimensional Gaussian distribution, they estimated the mean vector and covariance matrix of each of these classes (i.e. the letters).

In order to simulate the "real life" situation and noting that the design sample size for each class was quite large, the above estimated parameters for each class were considered as the true parameters of an 8-dimensional Gaussian distribution. For the rest of the paper these parameters will be consistently refered to as the true parameters. To simulate the "real life" situation, however, these distributions are assumed to be unknown to the classifier. Consequently in order to apply the classification rule, it is required to obtain estimates for these parameters for use in the classification process. This can be achieved by generating a number of training vectors according to each class distribution.

Since only two-class problems are investigated, the four classes were considered in a pairwise fashion. In order to simplify the results, it was decided to consider only two of these pairs of classes; the least and most separable pairs. To choose these two pairs the following steps were carried out. Each pair of covariance matrices were first simultaneously diagonalized to simplify the succeeding calculations. The features of each pair were then processed and ordered according to their effectiveness in discriminating between the two classes in that pair. The probability of correct classification was chosen as the measure of effectiveness for the different features. This choice is prefered to some "distance" measures since none of the known distance measures (e.g. the Divergence or the Bhattacharyya distance) is uniquely or monotonically related to the probability of correct classification. Thus the best feature, according to the above measure, was selected. Then considering this selected feature with each of the remaining features, the second best was thus selected. This process was then repeated until all features were considered. Note that a selected subset of some specified size is not necessarily the best subset of this size. From this analysis the least separable pair of classes was chosen to be that pair which resulted the smallest probability of correct classification, for only one feature. The two letters comprising this pair were the letters B and D which intuitive-

ly seems reasonable. Similarily the most separable pair was defined. The two letters comprising this pair were the letters C and D which again seem to agree with intuition. The above two pairs will be consistently refered to as pair B-D and pair C-D respectively.

For pair B-D a nonsingular linear transformation was applied to the observations such that the covariance matrix for the letter B is the identity matrix and the covariance matrix for the letter D (Σ_D) is diagonal. The elements of Σ_D and the absolute value of the transformed difference-of-mean vector μ_D are shown, ordered as explained before, in Table 1. The corresponding parameters for pair C-D are shown in Table 2. Note that in both tables the numbers in brackets represent the original indexes of the features as given by Marill and Green.

The purpose now is to study the effect on the probability of correct classification of using estimated class distributions in the classification process. To achieve this purpose it is assumed that the true distributions are not known by the classifier but a number of training vectors K are generated according to each of the class distributions. These vectors are then used to estimate the true parameters to be used in classifying unlabelled vectors. Two types of experiments for studying the effect of dimensionality on P_{cr} have been performed on the data shown in Tables 1 and 2. In the first experiment, a fixed number of vectors was used for parameter estimation for each dimensionality N. This situation simulates the "real life" situation in which a fixed number of design vectors are available with the freedom of choosing the number of features to be used. In the second experiment the number of training vectors is assumed to be functionally related to the dimensionality N as follows

$$K = \alpha N,$$

where α is an integer constant larger than one.

In the two experiments mentioned above the probability of correct classification for a particular design sample is a random variable simply because this sample is randomly generated. Since the intention is to study the effect of estimation in general regardless of the design sample used to estimate the parameters, it is necessary to average this probability over all possible design samples of the same size. The particular results presented were based on averaging the calculated value of P_{cr} over one hundred samples for each value of K and each value of N. This gave rise to relatively small statistical fluctuations in the results for all cases. Consequently the relationship between the average probability of correct classification $\bar{P}_{cr}$ and N was obtained with K used as a parameter in the first experiment and with α used as a parameter in the second experiment.

The results of experiment 1 are shown in Figs. 1a and 2a for pairs B-D and C-D respectively. The values of K considered were 5, 10, 20, 40 and infinity (known parameters case). Since for pair C-D the values of $\bar{P}_{cr}$ for k = 20 or 40 were very close to the known parameters case, these cases are not shown on the graphs. For experiment 2 the results are shown in Figs. 1b and 2b for the two pairs. The values of α considered were 2, 5, 10 and infinity. Again for pair C-D the case α=10 was not plotted for the same reason mentioned above. To give a rough idea about the efficiency of the proposed technique, it is important to mention that it took less than 90 seconds on the Dec system 2050 machine to calculate the average probability of correct classification (over 100 samples) for any K and for any N between 1 and 8.

From the above results it can be concluded that the proposed technique is quite efficient. From the results of experiment 1 it is noted, as expected, that peaking is always encountered in the $\bar{P}_{cr}$ versus N relationship for any finite sample size. For the second experiment, however, since K is increasing with N, peaking generally does not occur. This latter result suggests that in order to obtain good estimates for the covariance matrices, and consequently minimize the effect of estimation on P_{cr}, it is necessary that at least 10N design vectors are available from each class. If the number of design vectors is proportional to N (i.e. equal to αN where α is a constant larger than one but not necessarily larger than 10), then peaking might be avoided but still the performance could be far from the case of infinite sample size.

V. SUMMARY

In this paper a technique has been developed to numerically calculate, as opposed to estimating, the true probability of correct classification for the two-class multivariate Gaussian problem wherein classification is based on estimated parameters. This technique in essence represents an extension of Fukunaga's technique which can be directly used only if classification is based on the true distributions.

In the proposed technique, as in Fukunaga's technique for the special case mentioned above, each class probability of correct classification is expressed in terms of a univariate integration of a simple and reasonably well behaved function. Thus the true overall probability of correct classification is obtained by performing two univariate integrations that can be numerically evaluated.

The proposed technique has been successfully implemented and it has proven to be quite efficient and reasonably simple. Real data have been used to demonstrate the applicability and efficiency of the technique. From the results obtained on the data, it has been noted that unless enough design vectors per feature are available to estimate the parameters, peaking is always encountered especially if the features are ordered according to their classification power.

VI. ACKNOWLEDGEMENT

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada (Grant A3622), the University of Saskatchewan and the Energy, Mines and Resources (agreement 107-5-79) for continued support while carrying out this research.

VII. REFERENCES

1. K. Fukunaga, Introduction to statistical pattern recognition. New York: Academic, 1972.

2. K. Fukunaga and T.F. Krile, "Calculation of Bayes' recognition error for two multivariate Gaussian distributions", IEEE Trans. on Comput., Vol. C-18, No. 3, March 1969, pp. 220-229.

3. C.R. Hallum, "Feature selection via an upper bound (to any degree tightness) on probability of misclassification", proceeding of 1st symposium on machine processing of remotely sensed data, October 16-18, 1973, pp. 3B-13 - 3B-26.

4. G. Mobasseri and C.D. McGillem, "Multiclass Bayes error estimation by a feature space sampling technique", IEEE Trans. on systems, Man and Cybernetics, Vol. SMC-9, No. 10, October 1979.

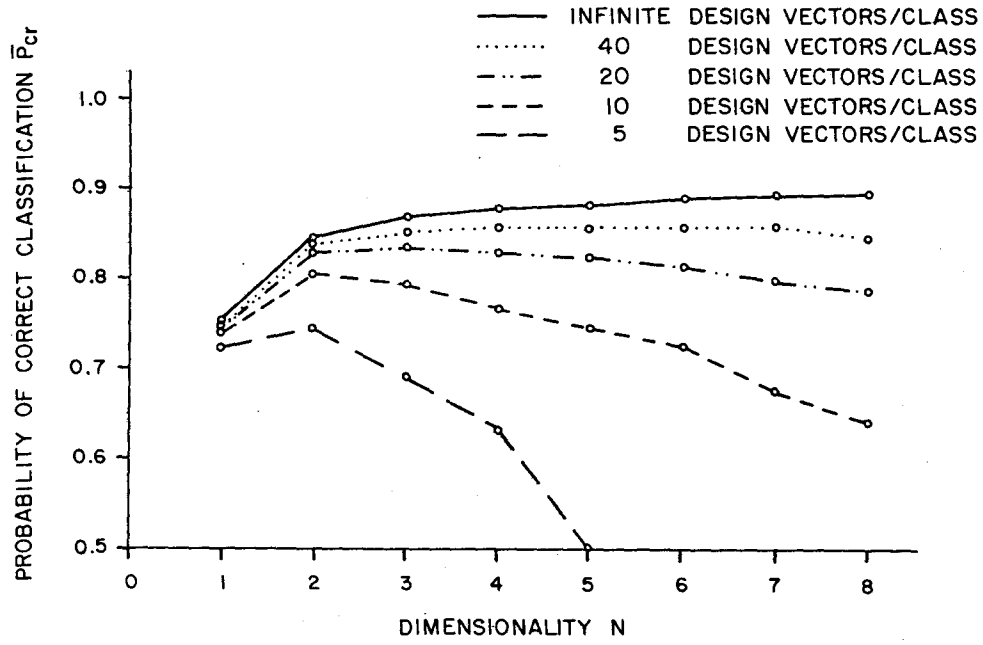
5. T. Marill and D.M. Green, "On the effectiveness of receptors in recognition systems", IEEE Trans. Information Theory, Vol. IT-9, pp. 11-17, January 1963.

Table 1. Ordered Transformed Parameters for Pair B-D. The covariance matrix for B is the identity matrix, the covariance matrix for D (Σ_D) is diagonal and μ_D is the transformed difference-of-mean vector.

Features	1(5)	2(7)	3(4)	4(8)	5(6)	6(1)	7(2)	8(3)
Elements of Σ_D	0.226	0.110	0.590	1.227	1.084	2.384	1.800	1.422
Elements of $\lvert\mu_D\rvert$	0.873	0.255	0.684	0.783	0.641	0.316	0.061	0.010

Table 2. Ordered Transformed Parameters for Pair C-D. The covariance matrix for C is the identity matrix, the covariance matrix for D (Σ_D) is diagonal and μ_D is the transformed difference-of-mean vector.

Features	1(6)	2(8)	3(7)	4(1)	5(3)	6(5)	7(2)	8(4)
Elements of Σ_D	0.203	0.012	0.025	17.277	5.158	1.244	7.839	1.616
Elements of $\lvert\mu_D\rvert$	2.569	0.072	0.414	5.869	3.716	1.369	0.190	0.506



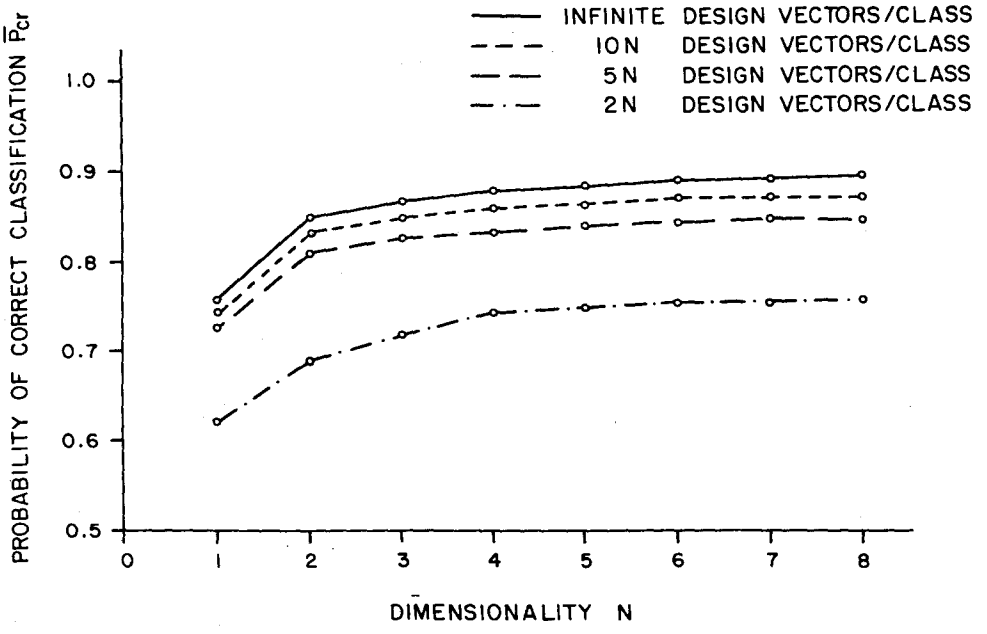
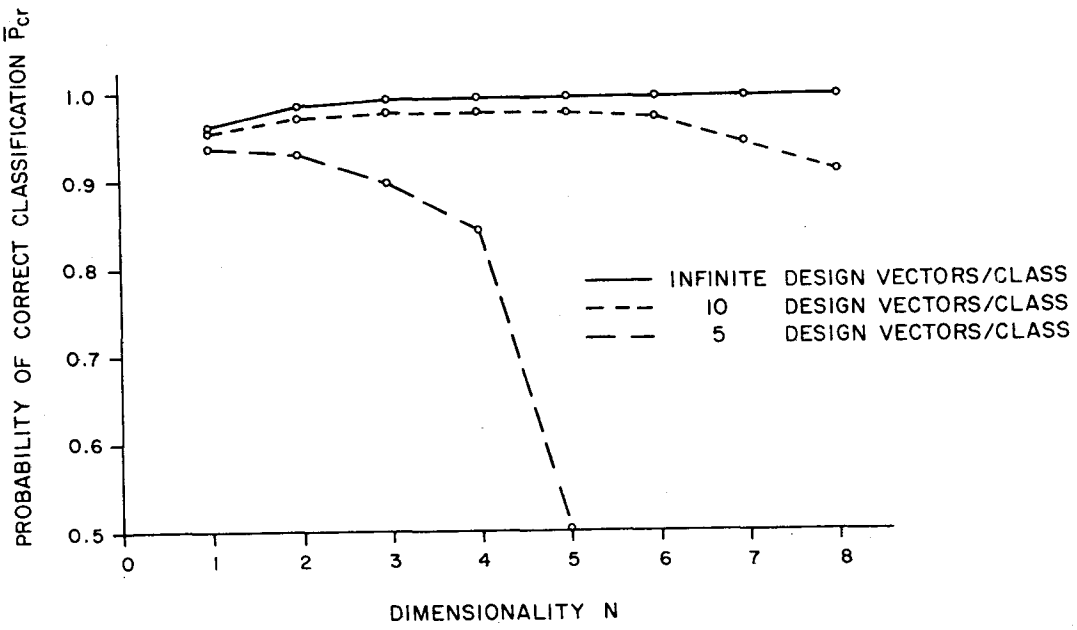


Figure 1. Average Classification Accuracy (Pair B-D).
(a) Fixed number of design vectors. (b) Number of design vectors proportional to dimensionality.



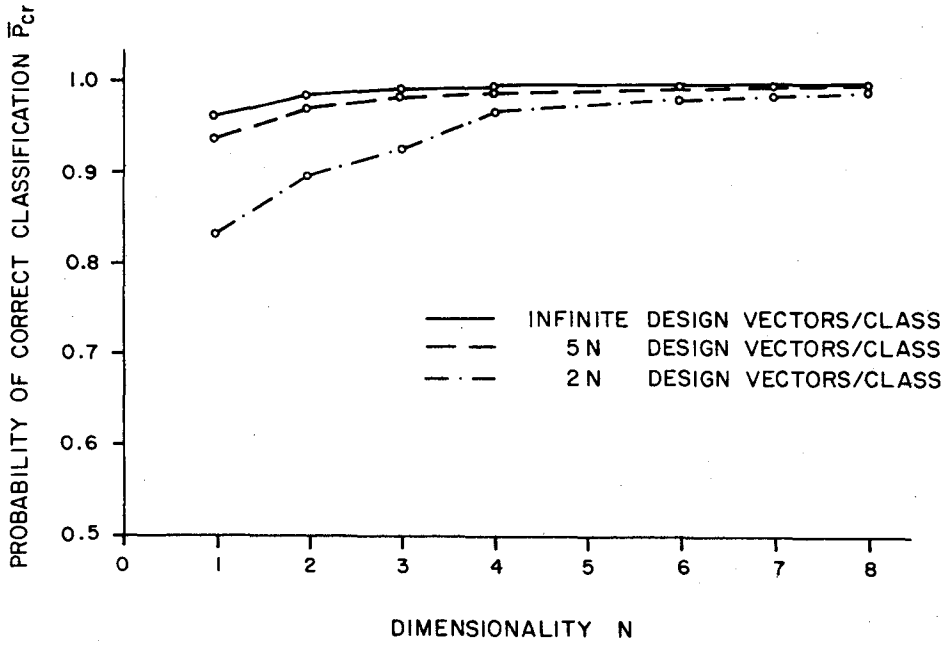


Figure 2. Average Classification Accuracy (Pair C-D).
(a) Fixed number of design vectors. (b) Number of design vectors proportional to dimensionality.